OPTICAL TEXT RECOGNITION: BASIC PROCEDURES AND CURRENT STATE

Danijel Radošević,
University of Zagreb,
Faculty of Organization and Informatics,
42000 Varaždin, Pavlinska 2
E-mail : darados@foi.hr

Slavko Vidović
University of Zagreb,
Faculty of Organization and Informatics,
42000 Varaždin, Pavlinska 2
E-mail : slvidovi@foi.hr

Abstract: *The survey of today's state of tools for optical text recognition is given in this scientific paper. Tools for processing handwritten symbols still did not enter in wide usage except in some specific cases such as hand-held computer. In the context of this scientific paper, given solutions were used in program "Handwritten Symbol Recognition". Today, on the other hand, tools for printed text recognition are already in wide usage. In the context of this scientific paper, tests of speed and accuracy of the recognition had been carried out for few today's popular commercial tools.*

Keywords: *Text recognition, handwritten symbol recognition, recognition of the patterns, OCR, algorithm.*

1. INTRODUCTION

Optical text recognition is a part of general theory for recognition of patterns and to these days it is probably the most developed part of it. Causes for this are enormous amounts of different written material that are necessary to input to the computer in order to be elaborated, but on the other hand those problems were a bit easier to solve in satisfying way than, for example the recognition of pictures. Yet, even here things are not simple, as for example the handwritten symbol recognition from the outside media (mostly paper work) is far more complicated than the procedure of recognizing printed text. Considering all these, tools for recognizing printed text are lately in commercial offer. Those simpler samples are delivered with scanner, but tools for recognizing handwritten symbols are mostly in developing phase (except some specific application such as entering commands at hand-held calculator because there are limited number of keys on the keyboard).

What can we expect from tools for text recognition?

At the first place, they should liberate us from hard working and long lasting typing in the text to the computer, if that text already exists in the paper media. For example, if typist types 150 characters per minute, she will need approximately 13 minutes per one printed page of 2000 characters. In case of using the program for text recognition that time will come to 5 - 10 seconds for exchanging page, approximately 15 - 30 seconds (depending on scanner) for scanning the text and approximately 5 seconds for text recognition by program recognition. It will be 30 - 45 seconds mostly. Calculation can only be deranged by question of accuracy of recognition. It depends on several factors and the most important among them is quality of scanned text, scanning quality and quality of program for text recognition. The subject of this scientific paper is quality of program for text recognition, according to that text quality for scanning is mostly given in advance and scanning quality has reached the satisfying level.

2. SYSTEMS FOR RECOGNIZING PATTERNS

Systems for recognizing patterns (according to [7, 203]) has in common that by various digital entities try to join up codes, that is to say (i.e.) to enable processing of information on the computer. There are many systems, which are different according to the type of patterns they are processing as well as by the procedure of processing. According to type of patterns, systems for recognizing patterns can be divided on:

a) Systems for recognizing sound patterns (mostly pronounced words) which are divided on :
   - systems that depend on speaker and
   - systems that don't depend on speaker

Further on:

- systems with limited range of words (only for giving commands from limited group)
- systems for recognizing free speech (enable dictating)

b) Systems for recognizing visual patterns :
   - systems for recognizing printed text
   - systems for recognizing handwritten text
   - systems for recognizing pictures / photos (e.g. recognizing persons)
   - systems for recognizing pictures in movement
   - systems for recognizing three-dimensional patterns

Beside mentioned systems, there are some specialized systems, which represent subspecies of previously mentioned. For example: systems for recognizing linear code (barcode), systems for recognizing numbers etc.

c) Other systems for recognizing patterns:
   - systems for processing time series, e.g. electrocardiogram

Such systems are usually based on time-series analyses.

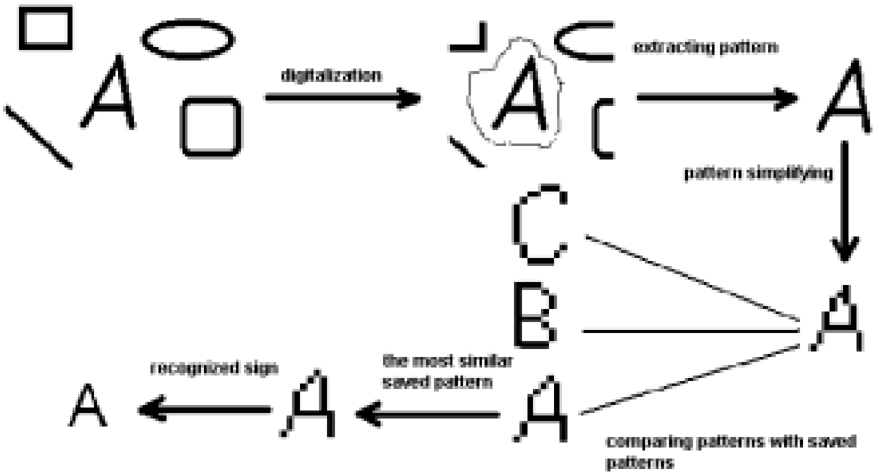
Up today, the widest usages have some specialized systems for recognizing linear code, which represents standard for marking the product. Wider systems, such as those for recognizing printed text are in broad usage in last few years, which enables computer to read information out of media that is adjusted to men.

3. GENERAL PROCEDURE FOR RECOGNIZING PATTERNS

In most cases, procedure for recognizing patterns is consist of next few phases:

1) *Phase for digitalization of patterns.* In this phase, the pattern is inputted to the computer by some entering devices, in most cases it is optical scanner, digital camera or microphone.
2) *Extracting pattern from the surrounding.* During digitalization, together with pattern of recognition, in most cases information from surrounding patterns or hum are inputted to the computer. That's why is necessary to establish what belongs to pattern recognition.
3) *Phase of simplification of pattern.* Usually, digitized pattern has great number of different qualities and it's impossible to process all of them. It's not necessary anyway, because all qualities are not significant for recognition. That's why in this phase, the number of qualities is reduced, and i.e. pattern is transformed into a shape suitable for recognition. According to [6.142], it's possible to:
   a) extract qualities of patterns
   b) record simplified picture of pattern

Recognizing patterns phase where they are compared to elements of basic patterns. In that phase, similarity or deviation of patterns for recognition of each pattern is confirmed. In that way, the most similar pattern to which is joined suitable sign are found.


digitalization
extracting pattern
pattern simplifying
comparing patterns with saved patterns
the most similar saved pattern
recognized sign

Picture 1.: General procedure for recognizing patterns

4. SYSTEMS FOR HANDWRITTEN SYMBOL RECOGNITION

It is believed that handwritten symbol recognition is one of complicated problems in the field of pattern recognition. However, different variants of this problem are possible. Those are mutually different in complex and in various approaches in solving and so we may differ:

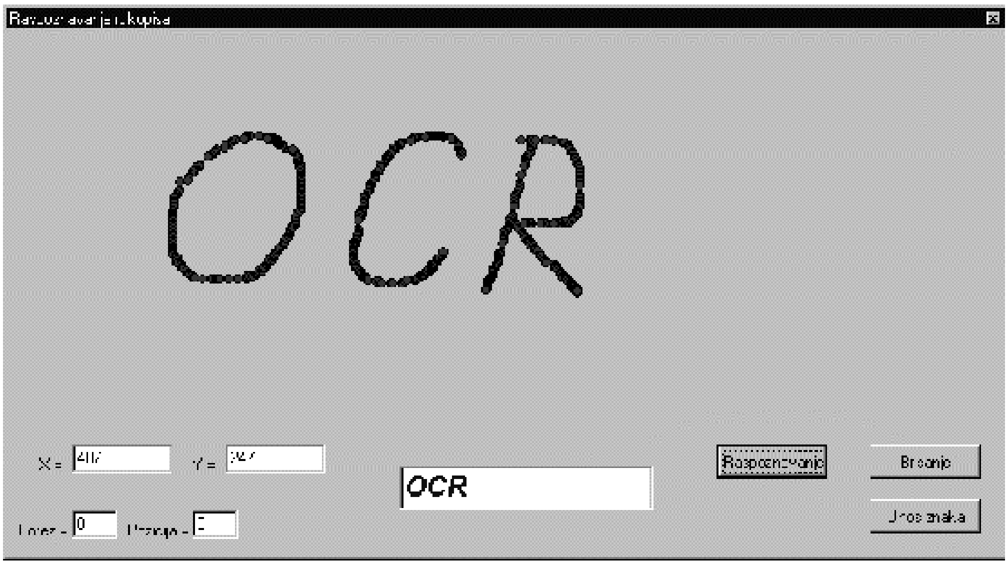
a) recognition of hand written characters that are put into the computer by some of the entering computer devices (e. g. mouse or graphic slab). This problem is relatively the simplest because digitalization of the text is managed simultaneously with the entrance so it's possible to follow movement of the hand and analyzed it later on.
b) recognition of, previously recorded characters from paper or some other media during which simultaneous digitalization with entrance is not possible but scanning is needed. In that case, during entrance it's not possible to follow movement of the hand and that makes problem more complex. Today, there are notable projects in that field such as CEDAR [2, 1] project.

Complexity of the system is bigger if the independent of the user is demanded. That is to say, if recognition of handwriting of the unknown user is needed (different from that who made entrance for pattern recognition).

4.1. The "HANDWRITTEN SYMBOL RECOGNITION" PROGRAM

"Handwritten symbol recognition" program demonstrates one of possible approaches to solving the problem (recognition of hand recorded characters into computer by the assistance of entering computer devices). Program enables following:

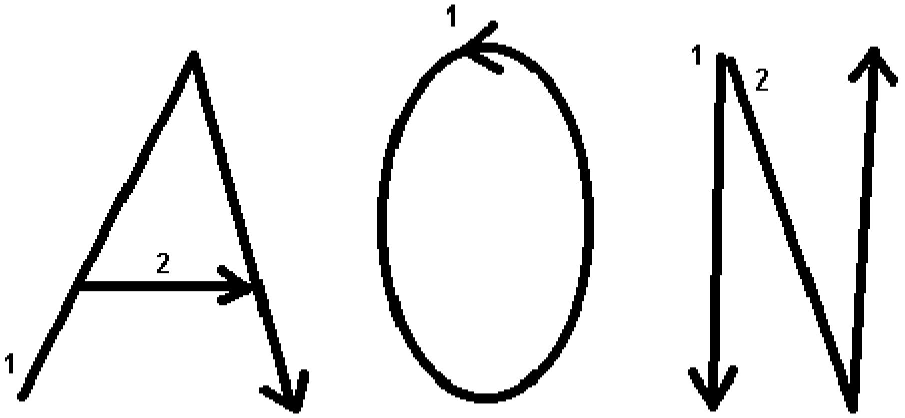
- entrance of particular characters or words into the computer and their recognition
- "learning", i.e. entrance of new samples into computer because of accurate recognition.



Picture 2.: Look of the screen of "Handwritten Symbol Recognition"

4.1.1. WAY OF WORK OF "HANDWRITTEN SYMBOL RECOGNITION" PROGRAM

"Handwritten symbol recognition" program enables the user input of characters by mouse or other compatible devices such as, for e.g. graphical slab. It is possible to write out character in one or more moves (picture 3).



Picture 3.: Different moves at printing characters (numbers are defining ordinal number of strokes)

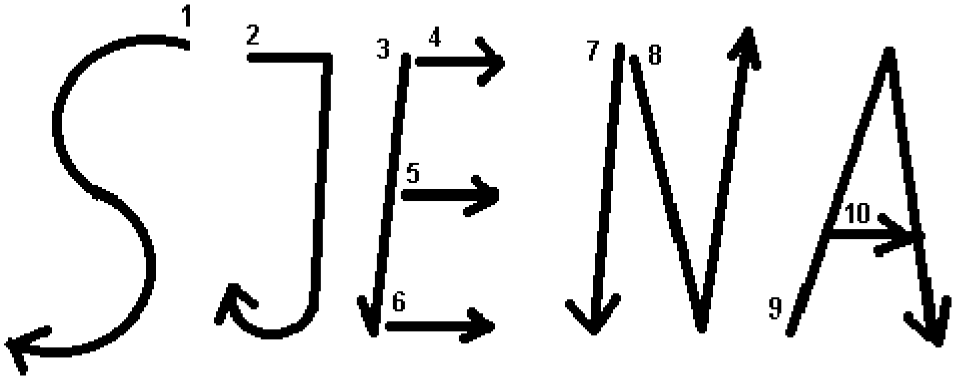
4.1.1.1. RECOGNITION PROCEDURE AT "HANDWRITTEN SYMBOL RECOGNITION" PROGRAM

Recognition procedure can be divided in few phases:

1) record of mouse coordinates (X, Y) in row, which make individual moves in adequate series (A_1, A_2…An),
2) extracting of characters, i.e. establishing which moves belong to certain character
3) standardization of characters (transformation into shape which is suitable for recognition) and
4) recognition of characters contrasting isolated characters with samples

4.1.1.2. RECORD OF MOUSE COORDINATES IN A ROW

One movement is made by all positions of cursor while the left key on mouse was pressed down. That kind of positions are recorded in suitable row of coordinates (A_1, A_2… A_n - picture 4)



Picture 4.: Individual strokes are recorded into adequate series

4.1.1.3. EXTRACTION OF ISOLATED CHARACTERS

In extraction of isolated characters it's necessary to determine which moves are entering in particular character. The procedure is based on calculation of the smallest distance between dots in two movements. Presumption is that two movements belong to the same character if two closest dots are very near to them, i.e. if their distance is smaller than limited. For example, movements 7 and 8 (picture 4) belong to same character and movements 8 and 9 belong to different ones. In case of movements 4, 5 and 6, which don't have close dots, these close dots exist in movement 3, which mutually "links" the whole character.

By the procedure of extraction of characters, from existed row of coordinates (A_1, A_2… A_n) new rows (B_1, B_2 … B_n) are formed. These rows have coordinates of all dots that belong to particular character (instead of movement, such as in a row A). In a sample of picture 4, it would look like this:

$B_1= A_1$
$B_2 = A_2$
$B_3 = A_3+A_4+A_5+A_6$
$B_4 =A_1+A_7+A_8$
$B_5 = A_1+A_9+A_{10}$

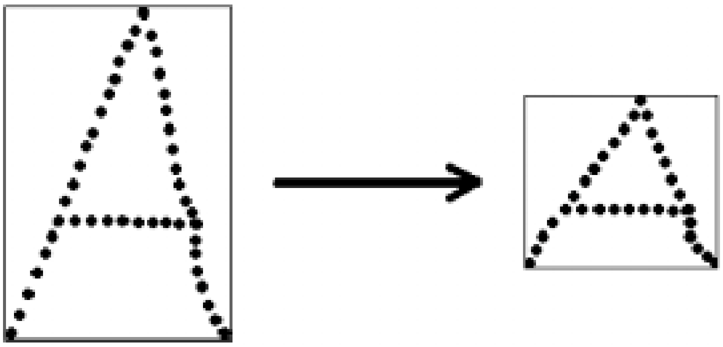
The procedure for extracting of particular characters is valuable by the presumption that the user inputted characters in a row, i.e. in a case he inputted them from right to left so they will be recognized.

Another presumption is that all parts of character are mutually close enough, and at the same time far enough from neighboring characters. For example, the problem may come out at letter "Š" where "clasp" must not be too far from "S".

4.1.1.4. STANDARDIZATION OF CHARACTERS

Standardization of characters consists of chain of transformations by which the influence of various ways of writing each character is brought to the smallest measure, to recognition on one hand and on the other make the character comparable to existed samples by which it is compared to.

In the example of " Handwritten symbol recognition " program standardization is managed considering coordinates of each dot of which the character consists of and decreasing number of dots to given number (p), which is equal to that as in pattern comparison (picture 5). The presumption is that character consists of at least as many dots as it is in pattern comparison.



Picture 5.: Standardization of character onto given size and number of dots

The result of standardization is new row of dots coordinates (C_1, C_2, … Cp), which is comparable to adequate row of samples.

4.1.1.5. RECOGNITION OF CHARACTERS

After the character is separated from the surrounding and transformed into the shape suitable for recognition by standardization it is necessary to compare it to characters which exists in data base .The procedure is as following:

a) it is necessary to compare the row of coordinates of dots the character for recognition consists of (C_1, C_2, …Cp), to suitable row of each pattern (D_1, D_2, …Dp):
    - in calculating distances of pairs of dots (C_1- D_1, C_2- D_2… Cp-Dp) Pythagorean theorem is used:

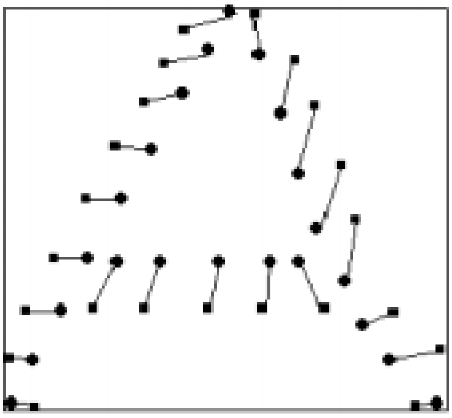
$$d_1 = \sqrt{(C_1 \cdot x - D_1 \cdot x)^2 + (C_1 \cdot y - D_1 \cdot y)^2}$$

$$d_p = \sqrt{(C_p \cdot x - D_p \cdot x)^2 + (C_p \cdot y - D_p \cdot y)^2}$$

    - an approximate distance of pair of dots is calculated:

$$\overline{d} = \underset{q=p}{\overset{q=1}{}} d_q / p$$

b) it is considered that the most similar sample is the one that gives the least distance of pair of dots for given character (picture 6)



- ● the dot of character to be recognized
- ■ the dot of the sample

Picture 6.: the recognition of characters in comparison to samples

4.2. THE PROCEDURE OF RECOGNITION OF PRINTED TEXT

The procedure of recognizing the text can be according to [4, 18] divided in three phases:

1. getting the text bit-map
2. processing of text bit-map by the program for optical recognition of the text and
3. the processing of the text got in phase (2) in one of standard programs for processing the text

In the first phase the digitalization of the text recognition is used, mostly by scanning. In that way the processing of the text by optical recognition is enabled. In the second phase the bit-map of the picture is transformed into the text that is further on possible to process in some of programs for text processing. In the third phase the correction of left errors in the text is possible (because of recognition is not perfect) and the editing to the final form of the text is enabled, too.

4.2.1. PROCESSING OF THE TEXT BIT-MAP

Processing of the text bit-map is the central phase of the text processing in which all qualities are expressed, i.e. defects of the program for text recognition. In that phase the following procedures are included, according to [4,19]:

- analysis of the page, here blocks of text for recognition are established
- text recognition inside of isolated blocks,
  - extraction of particular characters in text bit-map, and
  - joining the particular code to the characters:
    a) recognition on the base of samples, and
    b) recognition on the base of quality of the shape
- composition of characters
- text composition

4.2.2. PROCESSING OF THE TEXT GOT BY THE RECOGNITION

After the processing of the text bit-map we got the text as the result of recognition. Such text, however still has quite errors. Typical example is switching the letters "I " and " I" because there are no difference between them in some fonts. Because of these, additional text processing is needed to solve the problem. Such processing is based on [3,7]:

- usage of the context. In that case conclusion can be as following:
    - if the capital letter is in the word in which all others are written in small (and it is not the first one) than it is wrong:
        - if the wrong letter is “ I “ (big I) than “ l “ (small L) is correct one
        - if the wrong letter is “ O “ (big O) than (little O) is correct one
    - if the letter is in the surrounding of numbers (left and right characters are numbers) than it is wrong:
        - if character “ l “ (small L) or character “ I “ (big I) are wrong, than “ 1 “ (one) is correct
        - -if character “ O “ (big O) is wrong, than “ 0 “ (zero) is correct
- in the usage of program for spelling checker such processing is interactive i.e. the work of the man is required. In that case the computer notifies “ suspicious “ words and user decides about the replacement. Spelling checker can be a part of text recognition program (which is rare) or it can be used in the part of text processing
- The hand editing of the text, can be as in the previous used within the program of text recognition (if it has such possibility, as they mostly have) or within the program of text processing
- Editing of the text using the text-processing program. In that phase errors of the text has already been corrected and editing of the text considering fonts, styles, arrangement of passages and so on, are yet to be done, what is not offered by the editor, which exist within the character recognition program

5. CHARACTERISTICS OF LEADING COMMERCIAL TOOLS FOR TEXT RECOGNITION

5.1. OMNIPAGE

OmniPage is the product of an American firm Caere, which offers several various tools for optical text recognition. In the moment of making this seminar the latest product of that firm is OmniPage Pro 8.0, and it is offered for Windows 95/98/NT.

The basic characteristics of it, according to [1,1] are:

- great accuracy, 99% for documents written on laser printer
- better recognition at, for computer, hard readable texts:
    - text in italics (e.g. if it is scanned by hand scanner)
    - photocopies of low quality
    - extremely tiny letters
    - inverse text (white letters on the black base)
    - more lingual text
- it keeps previous order of the page, e.g. paragraphs, form (shape) of the page, attributes of the fonts, slabs, photos
- notifies "suspicious" characters and propose corrections
- enables parallel correction of errors on recognized pages and further recognition of the text
- integration with Microsoft Office 95 and 97. Enables recognition of the text directly from Word or Excel. It shares the same vocabulary with these tools, also.
- possibility of storing documents in various forms, including HTML
- recognize specific signs of 11 languages

Additional possibilities are:

- 3D OCR – use scanning in half tones, because of easier recognizing text on colored pages

- Language Analyst – use dictionary and linguistic information for improving accuracy of recognition
- AccuPage 2.0 – enables better quality of recognition when scanners from Hewlett-Packard's ScanJet series are used
- Schedule OCR – additional starting of recognition in given time of a day

5.2 PERCEIVE PERSONAL

Perceive Personal was made by an American firm OCRON, Inc. It has been delivered recently with Qtronix hand scanners and it belongs to a group of simple ones, i.e. modest optical tools by its possibilities, for recognizing text.

Version 2.0 of this tool was made for operative systems of Windows 3.1 and it may be used on modern versions (95/98/NT).

It has the following characteristics:

- It may read documents from the scanners which work with TWAIN drive program, or from files in graphic forms TIFF and PCX
- Documents can be more lingual, or written in one of 11 languages but it doesn't recognize our specific signs (Č, Ć, Đ, Š and Ž)
- It has its own editor for correction errors on recognized text
- It can store recognized text in forms ASCII and ANSI, in variants with or without leaving the sign for the end of the paragraph
- It can recognize fonts on its own but can't recognize extremely small letters, or extremely big ones
- it can relatively bad solve problems of pictures, slabs, text scanned in italics, and text printed on registered printer
- it can't process analyze of the page

5.3 READIRIS

Readiris is a product of Image Recognition Integrated Systems firm, and it is delivered with Primax scanners. In different versions it exist for operative systems of Windows 3.x/95/98/NT.

It has the following characteristics:

- it can read documents from scanners that work by TWAIN drive program, or from files of graphic forms TIFF (compressed and non compressed), PCX and MSP (MS-Paint) form.
- It can analyze a page with possibility of changes from user (picture xjjj)
- Enables interactive "learning" i.e. defining of unknown (signs) characters
- Documents can be in one of 21 languages, Croatian as well, i.e. it can recognize our specific signs (Č, Ć, Đ, Š and Ž)
- It can store recognized text in one of 13 output forms, including those for the most popular programs of text processing (Word, Word Perfect, WordStar) and slabs form of table calculator Excel
- It can automatically extract pictures from the text and stores it into special files
- Recognize slabs
- It can recognize fonts on his own, but for extremely small ones adequate options must be installed
- There is an option for recognizing texts printed on matrix printer (but in that case results of recognizing text are relatively small)

- It doesn't contain his own editor for correction of the errors, but it can be called out of program Word, WordPerfect or Excel where the results of recognition are received
- It doesn't contain vocabulary for spelling checker of recognized text, but part of the errors can automatically be corrected on the bases of linguistic context

5.4. RECOGNITA

Under the title of Recognita, homonymous Hungarian firm, nowadays more products on the market are offered, among which the most famous are [5,1]:

- Recognita Plus, at the moment in version 4.0, is at the time considered the best tool for text recognition which is intended for PC
- Recognita Select, which is delivered together with HP scanners and
- Recognita Development Toolkit, contains program modules for development of applications which use OCR

Recognita Plus is professional tool for optical recognition of characters, which beside exceptional accuracy of recognition, offers following:

- Intelligent analysis of the page
- Recognition of characters of 107 various languages
- Recognition of the slab
- Enables group recognition of pages
- Independence of font and size of text are achieved
- Sort out pictures in color as special files
- Enables 100 various output forms of text
- Contains own editor for correcting errors in recognized text
- Keeps previous schedule of the page and writing styles
- Enables recognition of telefax, barcode, text written in matrix printer and hand written numbers

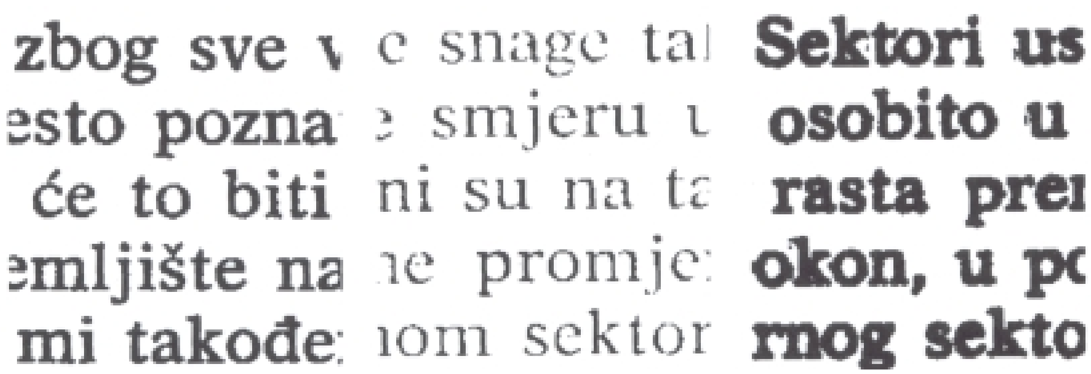
Recognita Development Toolkit contains program modules for development of applications that use OCR. Modules can be called out from various program languages. In that way great flexibility of optical recognition of characters is achieved as well as its better integration into information system.

6. COMPARISON OF FEW TOOLS FOR TEXT RECOGNITION

6.1. SPEED AND ACCURACY OF RECOGNITION TESTS

Tests have been carried out on three texts (picture 7), which were of various qualities, and speed and accuracy of recognition have been measured, too:

- Text No. 1: Optimal possible quality of given text (the least possible number of "stick up" characters, in other words, incomplete characters)
- Text No. 2: text scanned by too high value of limit at scanning so that letters are pretty much "crackled", i.e. there are many incomplete characters
- Text No. 3: text scanned by too low value of limit at scanning, so that letters seem "bold" with great number of cases of "stick up" letters



zbog sve ∖ c snage ta∖ Sektori us
əsto pozna ∶ smjeru ι osobito u
će to biti ni su na tε rasta prei
əmljište na ıe promje∶ okon, u pc
mi takođe∶ ıom sektor rnog sekto

Text No. 1 Text No. 2 Text No. 3

Picture 7.: Text recognition patterns

Texts are of one printed page size, each contain about 2000 characters and are scanned in 300 DPI resolutions, in one-colored (line-art) way. Table scanner A4 form is used. For tests Pentium computer is used at working fact of 133 Mhz. Results mentioned in table 1 have been achieved by testing.

Program	**Text No. 1**		**Text No. 2**		**Text No. 3**		Average speed (c/s)	Average accuracy
	Chars / sec	Accuracy	Chars / sec	Accuracy	Chars / sec	Accuracy		
Cuneiform 2.0	467	97,00 %	280	94,00 %	138	90,50%	295,00 %	93,83
Megaznak	74	83,50 %	65	50,50 %	85	61,50 %	74,67 %	65,17
Perceive Personal 2.0	519	94,50 %	448	93,00 %	245	80,50 %	404,00 %	89,33
Readlris 3.50g	240	99,00 %	172	92,50 %	122	87,50 %	178,00 %	93,00
Recognita Plus 2.0	812	98,50 %	576	89,50 %	106	89,00 %	498,00 %	92,33
Recognita Select 2.0	718	97,50 %	579	97,50 %	83	90,00 %	460,00 %	95,00
Average	472	95,00 %	353	86,17 %	130	83,17 %	318 %	88,11

Table 1.: The results of text recognition for several programs for recognition of text

6.2. COMPARISON OF PROGRAM POSSIBILITIES OF TEXT RECOGNITION

[redacted]	Programs	Cuneiform 2.0	Megaznak	Perceive Personal 2.0	Readlris 3.50g	Recognita Plus 2.0	Recognita Select 2.0
Characteristics	[redacted]						
Analysis of the page		YES	NO	NO	YES	YES	YES
Automatically recognition of fonts		YES	NO	YES	YES	YES	YES
Number of maintained languages		9	2	11	21	21	23
Maintenance for Croatian specific letters		NO	YES	NO	YES	YES	YES
Tables		YES	NO	YES	YES	YES	YES
Extraction of pictures		YES	NO	NO	YES	YES	YES
Supplemental processing		YES	YES	NO	YES	YES	YES
Possibility of "learning" new characters		YES	YES	NO	YES	YES	NO
Input forms of pictures		TIFF, PCX	PCX	TIFF, PCX	TIFF,PCX, MSP	TIFF, PCX	TIFF, PCX

Number of output text forms	5	1	5	13	80	35
Possibility of collective recognition	YES	YES	NO	NO	YES	NO
Calling up from text processor	YES	NO	NO	YES	NO	NO
Own editor form recognized text	YES	NO	YES	NO	YES	YES

Table 2.: Comparison of possibilities of several programs for optical text recognition

7. CONCLUSION

Nowadays, optical text recognition reached maturity on the field of recognition of printed text. Nowadays, it can be seen by that the programs of lower class are regularly delivered together with scanners, and professional tools are offered. By further development of such tools efforts are made as follows:

- Increase of accuracy for text recognition of lower quality
- Truthful reproduction of the page of recognized text, i.e. except characters itself, efforts are made to recognize and edit pages (columns, slabs, pictures) and styles and fonts
- Integration for recognizing text with other applications (text processing at the first place)

Recognition of handwritten symbols is in developing phase for now, and first commercial applications appeared in a last few years, mostly in very specific fields such as hand-held computer where they replace keyboard. In that field there are still many unsolved problems, so wider application on PC is expected in the period of next few years.

LITERATURE

[1] Caere Inc. (1998): *Caere scanner OCR software*, http://www.Caere.com, taken over on 4th of August

[2] CEDAR (1998): *CEDAR Home Page*, http://www.cedar.buffalo.edu, taken over on 29th of July.

[3] Radošević, D. (1990): *Recognition of two-dimensional forms and pronounced words by using similar algorithms*, Anthology of works Faculty of organization and informatics, No. 14, Varaždin.

[4] Radošević, D. (1996): *Basic procedures and problems of optical text recognition*, Anthology of works Faculty of organization and informatics, No. 21, Varaždin

[5] Recognita Corp. (1998): *Recognita Products*, http://www.recognita.com, taken over on 3rd of August

[6] Ružić, F. (1994): *Multimedia: integration of sound, pictures and text by help of PC*, Mozaik knjiga, Zagreb

[7] Wayner, P. (1993): *Optimal Character Recognition*. Byte 12/93.